SCIENCA REVUO de Internacia Scienca Asocio Esperantista (BEOGRAD, Jugoslavio)	El Vol. 21 n-ro 4/5 (84/85) 5. 9. 1970.

SEMANTIKO KAJ STATISTIKO DE RUSAJ SUBSTANTIVOJ

H. D. Maas (Universistato Saarbrücken, Germanujo)

1. Enkonduko.

Okupiĝante pri maŝina traduko (MT) oni frontas unue la problemon de la frazo-analizo, kies rezultoj estas la bazo por la transferaj kaj sintezaj operacioj, liverantaj la tradukitan tekston. [1]

La tasko de la analizo estas la malkovrado de la rilatoj inter unuopaj vortoj kaj inter vortosekvoj. Specialan atenton tie postulas la homografoj, t. e. gramatike plursignifaj vorformoj, kies traduko eblas nur post kiam iliaj gramatikaj karaktero kaj funkcio en aktuala tradukota frazo estas konataj. La ĝisnunaj analizaj algoritmoj baziĝas sur la scioj die la sintakso de la traktataj lingvoj.

Sufiĉe ofte oni ne sukcesas solvi tiajn formalajn plursignifojn, pro kio la maŝino miskonstatas la gramatikan strukturon de analizata frazo, kvankam aŭdanto aŭ leganto tuj povas kompreni ĝin. Do, fakte ne ĉiu nesolvebla malfacilaĵo estas tia por homo. Supozeble la homo intuicie disponas ne nur pri sintaksaj konoj, sed ankau povas utiligi semantikajn kriteriojn por la solvo de formalaj plursignifoj.

Studante la eblecojn de MT oni direktas sian specialan atenton pro tio al la demando, ĉu kaj kiel semantika analizo estas formaligebla kaj maŝine aplikebla. La renkontotaj malfaciloj estas jam konataj: Alie ol en la domajno de sintakso, kie la nocioj de regado aŭ valento, de vortoklaso ktp. povas esti formale difinitaj, kie oni posedas abstraktan modelon por la sintaksa strukturo de naturaj lingvoj en la formo de transformacia gramatiko, en la semantiko ne ekzistas pli-malpli akceptitaj »vortoklasoj» kaj »valentoj«, nek ĝenerala modelo priskribanta la semantikajn rilatojn de la vortoj interne de iu frazo Ĝis nun estas eĉ ne imageble, kiamaniere oni difinu semantikan klasifon cele al objektiva decido, kiuj lingvaj elementoj apartenas al unu aŭ alia semantika klaso aŭ kategorio. Evidente ne ekzistas rigora limo inter semantiko kaj sintakso, ĉar ofte la decido, ĉu iu fenomeno estas origine semantika aŭ sintaksa, estas malfacila; eble tia apartiganta limo tute ne ekzistas. Ni memorigas ĉi tie, ke kelkaj verboj regas lokajn aŭ direktajn adiektojn, ekz-e loĝi«: Oni ne povas diri nur »mi loĝas«, ĉar »loĝi« postulas lokan komplemen-

ton, do ĝusta estas »mi loĝas en dometo« aŭ »mi loĝas hejme«. Tiun ĉi fakton oni povas klarigi enkodukante la nocion de loka kaj direkta valento — tio estus sintaksa priskribo — aŭ parolante pri la semantika klaso de lokaj verboj aŭ verboj de movo kaj moviĝo.

Tamen, se oni eksplikas la ekziston de loka adjekto per enkonduko de loka valento, oni devas atentigi, ke ne ĉiuj substantivoj povas aperi en la »loka kazo«. Per tio oni denove transiras al la semantiko: La substantivojn, kiuj povas aperi en la loka kazo, ni nomos »lokaj« au »lokaloj«, per kio ni difinis semantikan kategorion.

Nia ekzemplo de la loka kategorio certe estas tro simpligita kaj ne respondas en ĉiuj detaloj al la lingva realo. Oni povas konstrui lokan adjekton en la formo de prepozicia nominala grupo ne uzante pure lokan substantivon. La grupoj »en ĉambro«, »sur la strato«, »ĉe angulo« reprezentas lokajn adjektojn, la entenataj en ili substantivoj »ĉambro«, »strato«, »angulo« certe apartenas al la lokaloj; sed kvankam »en mia mano« estas loka adjekto, la vorto »mano«certe ne estas loka substantivo en la normala senco.

Serĉante objektivan metodan por klasifo de rusaj substantivoj, ni pensis unue pri tio, ĉu statistikaj donitaĵoj povas kontribui tiudirekte. Ni eliris de la supozo, ke por substantivoj el certaj semantikaj klasoj certaj sintaksaj konstruoj estas aparte tipaj. Substantivo, kiu indikas personon verŝajne plej ofte aperas en nominativo. Por la rusaj substantivoj ni atendas, ke lokalo aperas tre ofte en prepozitivo, ĉar tiu ĉi kazo estas regata preskaŭ sole de lokaj prepozicioj.

Ni interesiĝis nun pri tio, ĉu semantike parencaj substantivoj posedas ankaŭ kompareblajn kazo-oftecojn. Tiun ĉi esploron ni efektivigas por la rusa lingavo, kie ni disponas pri la statistika vortaro de Steinfeldt [2], en kiu estas listigitaj la plej oftaj substantivoj kun aldono de la ofteco de la ses kazoj kaj la du nombroj. El tiuj donitaĵoj ni elkalkulis por ĉiu unuopa vorto, kun kiel granda relativa ofteco ĝi aperis nominative, genitive, dative, akuzative, instrumentale kaj prepozitive en la studitaj tekstoj.

Nia pruvota hipotezo estas, ke semantike kompareblaj vortoj havas similajn probablecojn por la apero de la ses kazoj. Se tiu ĉi supozo pruviĝas kaj se la nocio »simila« koncerne probablojn povas esti precizigita, tiam ni disponos pri rimedo por konkludi sematikajn konstatojn el statistikaj donitaĵoj.

2. Universalaj kvantoj

Ni distingas du specojn da universalaĵoj: unue la kvantoj, kiuj restas konstantaj, se oni transiras de unu lingvo al alia; due la elementoj fiksaj en unu lingvo, ŝanĝiĝantaj sub influo de temo kaj stilo. Supre ni asertis, ke la probablo de la kazoj ĉe certa vorto estas tia universalo en la rusa lingvo. Same la kazo-oftecoj en iom pli grandaj tekstoj estas konstantaj. Universalaj por kelkaj lingvoj estas la probabloj de akuzativo kaj pluralo, se oni permesas al tiuj kvantoj ioman movoliberon. Parolante pri universaloj, oni kredeble devas ĉiam allasi tiajn movo-intervalojn. Por Esperanto ekz-e la meza longeco de vorto estas 1,859 silaboj, kion ni konsideras universala en tiu ĉi lingvo kaj kio apartigas ĝin de la angla lingvo, kie la valoro estas 1,351 [3]. Laŭ [3] la ofteco de unusilabaj vortoj en Esperanto estas 40,4%—en [4] estas indikita la nomro 35,06 % por la Andersenaj fabeloj, tradukitaj de Zamenhof. Ĉi tie ni ne insistas pri la diferenco, sed konstatas ke 40% estas universala cifero por la unusilabaĵoj. Por stilaj studoj tia devio tamen povas esti signifa.

En [2] estas sciigite, kun kioma ofteco aperis la ses rusaj kazoj. Car estis nombritaj pli ol cent mil substantivaj formoj, ni rajtas esti certaj, ke tiuj oftecoj estas universalaj probabloj. Komparante la kazo-oftecojn de unuopa substantivo kun tiuj ĉi mezumoj, oni povas decidi, ĉu la diferencoj estas signifaj aŭ malgravaj.

Por la posta uzo ni difinas kelkajn mallongigojn:

Dif. 2. 1.: La oftecon de nominativo ni nomas N, de genitivo G ,de dativo D, de akuzativo A, de instrumentalo I, de prepozitivo P.

El (2) ni citas la probablojn de la kazoj por

substantivoj	subst. kaj propraj nomoj
N=28,3%	N=33,6%
G=26,0%	G=24,6%
A=21,8%	A=19,5%
P=10,3%	P= 9,4%
I= 8,6%	I= 7,8%
D= 5,0%	D= 5,1%

Ni ekvidas, ke la diferencoj estas malgrandaj, sed kelkaj trajtoj estas jam aparte interesaj. La nominativo devas esti multe pli ofta ĉe propaj nomoj ol ĝenerale, kion ni supozis jam komence. La tabeloj montras ankorau plian interesan fenomenon: La ses kazoj povas esti dividitaj en du klasojn, ĉar evidente troviĝas fendo inter akuzativo — laŭrange la tria kun 20% — kaj prepozitivo — la kvara kun 10%. Atentinda ŝajnas al ni la fakto, ke plursignifeco de substantivaj formoj plej ofte okazas interne de ambaŭ klasoj N, G, A kaj P, I, D [5]. Ĉe substantivoj prisignantaj aferojn ĝenerale nominativo kaj akuzativo estas egalaj (krom ĉe femininoj), dum ĉe personosubstantivoj koincidas la genitiva kaj akuzativa formoj (krom ĉe femininoj).

Ni nin demandas, ĉu en aliaj kazoposedaj lingvoj la kazo-oftecoj similas al la ĉi-supre donitaj valoroj. Ni komparas kun sanskrito [1]:

nominativo	42,3%
akuzativo	27,4%
instrumentalo	9,1%
genitivo	8,2%
lokativo	6,4%
dativo	5,3%
ablativo	1,3%

Miriga estas la granda ofteco de nominativo kaj la relativa malofteco de genitivo. La koncernaj ciferoj por akuzativo, instrumentalo kaj dativo bone respondas al la probabloj konstatitaj por la rusa lingvo. Eĉ pli frapan akordon oni ekkonas inter la ofteco de singularo kaj pluralo. Kaj por sanskrito kaj por la rusa lingvo ekzakte 72% el ĉiuj substantivaj vortoformoj estas singularaj. Propra esploro de Esperanto tekstc kun 2183 vortoformoj[1]) liveris singularan oftecon de 68%. Laŭ testo de la statistika vortaro [7] kalkulis singularan oftecon de 75% por substantivoj kaj 73% por adjektivoj en la hispana lingvo. Por la vira kaj ina artikoloj ni konstatis 77% respektive. Do, ni konkludas, ke la probablo de singularo estas universala por tiuj ĉi lingvoj.

[1]) La teksto estis Rigveda. Ni citas laŭ Zipf [6].
[1]) Traktas pri »La internacia lingvo kiel esprimo kaj antaŭeniganto de universalismaj tendencoj« de I. Lapenna, en Elektitaj paroladoj kaj prelegoj, Rotterdam 1966.

La statistiko de la menciita Esperanto-teksto liveris oftecon de 23% por la akuzativo, do ankaŭ tiu kvanto ŝajnas universala.

Fine ni komparas tiujn ĉi ciferojn kun la kazo-oftecoj en la germana lingvo, kalkulitaj de Meier el tekstoj de sume 500.000 vortoformoj, el kiuj 150.000 estis substantivoj: [8]

N = 41,6% D = 24,9%
G = 9,4% A = 21,1%

Ankaŭ ĉi tie ni renkontas la konatan valoron por la akuzativo. La ofteco de la singularo estas rondcifere 78% — do, ankau bone akordiĝas kun la antaŭaj donitaĵoj.

Post tiu ĉi kompara superrigardo super aliaj lingvoj, ni turnu nian atenton denove al la rusa. Supre ni asertis, ke oni povas apartigi la rusajn kazojn en du klasojn surbaze de iliaj distribuoj, nome en la NGA-klason, kiu entenas nominativon, genitivon kaj akuzativon, kaj en la PID-klason kun la restantaj kazoj. Tiu ĉi apartigo eble ŝajnas triviala, ĉar ni kunigis ĝuste la tri plej oftajn kaj la tri plej maloftajn kazojn respektive en unu klason — tamen, ni montros ĉi-sube, ke la distribuo de la rusaj kazoj objektive igas nin akcepti tian duonigon. Ni bezonas novan difinon por tiu ĉi celo:

Dif. 2. 2: Se nia studata lingvo havas n deklinaciajn kazojn, nome k_1, k_2,- k_n, tiam k_i (X) estu la probablo por tio, ke la substantivo X aperas en la kazo k_i. Ordiginte tiujn ĉi probablojn laŭgrade, do

$$k_{i1}(X) > k_{i2}(X) > \ldots > k_{in}(X)$$

ni nomas la n-opon

$$(k_{i1}, k_{i2}, \ldots, k_{in})$$

kazovico de X.

Car la probabloj en la rigora senco neniam estas precize koneblaj, ni devas limiĝi je laŭofteca ordigo de la kazoj. Dum nia laborado ni konstatis, ke en sekvoj de neegalaĵoj de la supra speco preskau neniam valoras la signo de egaleco. Tial oni povas ĝenerale aldifini unu kazovicon al ĉiu substantivo. Ni opinias, ke tiuj ĉi kazovicoj por fiksa X estas universalaj, do sendependaj de temo kaj stilo. Tiuj faktoroj influas precipe la absolutan oftecon de X.

Se ni denove ĵetas rigardon sur la distribuon de la rusaj kazoj, ni povus konjekti, ke plej ofte aperas la »mezuma kazovico« NGAPID. Tio tamen ne ŝajnas esti vera: En nia studita materialo de la 450 plej oftaj substantivoj la kazovico NGAPID okazis nur dufoje. Kontraŭe aperis NGADIP dekfoje, GAPNID okfoje kaj NGIDAP deksesfoje. Sume ni konstatis 150 malsamajn kazovicojn.[1] Do, meznombre 3 substantivoj estis trovitaj por ĉiu kazovico.

Ni interesiĝis nun pri tio, kun kioma probablo la i-a kazo estas la j-a lau ofteco, konsiderante ĉiujn substantivojn. Estas do demandite, kiel ofte la nominativo (i = 1) estas la plej ofta (j = 1) kazo, kiel ofte la instrumentalo (i = 5) estas la plej malofta (j = 6) ktp. Kvankam ni ne scias, ĉu la maloftaj substantivoj havas aliajn kazodistribuojn kompare kun la pli oftaj, ni opinias, ke la rezultoj liveritaj de la statistiko de la menciitaj 450 vortoj estas sufiĉe instruaj.

Kiam ni signas per p_{ij} la probablon por tio, ke la i-a kazo staras sur j-a loko en iu kazovico, tiam valoras la sekvantaj ekvacioj:

[1] Maksimume ekzistas 6! = 720 malsamaj kazovicoj

$$p_{i1} + p_{i2} + \ldots + p_{i6} = 100\ \%$$

$$p_{1j} + p_{2j} + \ldots + p_{6j} = 100\ \%$$

La unua ekvacio eldiras simple, ke la i-a kazo devas aperi sur iu loko en la kazovico, dum la dua aludas la fakton, ke sur la j-a pozicio de kazovico devas troviĝi iu el la ses kazoj. La nombroj p_{ij} sekve formas matrikson, en kiu la kolonaj kaj la liniaj sumoj egalas 100% (=1). En la sekvantaj tabeloj ni atingas tiun ĉi valoron nur proksimume pro eraroj en nombrado kaj kalkulado.

Statistikinte la tutan materialon de 453 substantivoj, ni ekhavis la sekvantajn oftecojn p_{ij}

	N	G	D	A	I	P
1	32	28	1,5	26	3,5	8,5
2	24	29	1,8	26	5,7	13
3	21	21	7,0	23	15	13
4	16	15	16	18	23	14
5	3	4	32	7	30	22
6	2	2	41	1	25	28

Ekzemplo por la uzo de la tabelo:
$p_{43} = 23\%$: La probablo por tio, ke la akuzativo estas la tria kazo laŭ ofteco (ĉe iu arbitre elprenita substantivo) estas 23%.

Ni registris la rezultojn en diagramo, kies absciso estas $j \geq (1 \leq j \leq 6)$. dum la ordinataj valoroj estas p_{ij} Interligante per linio la punktojn apartenantajn al la sama kazo, oni ricevas ilustran bildon:

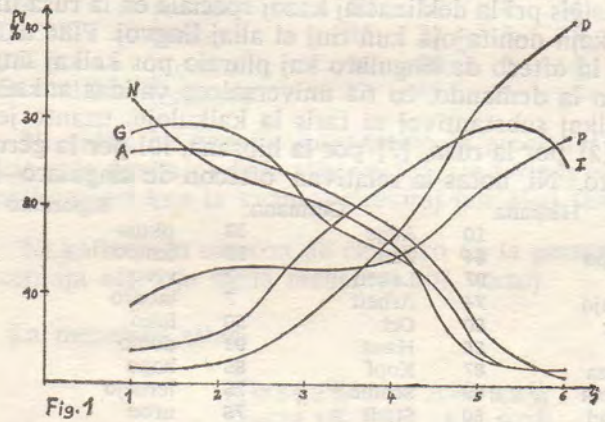


Fig. 1

La bildo montras tre klare, ke oni povas paroli pri du klasoj da kazoj en la rusa lingvo. La kurboj de la elementoj de la NGA-klaso estas tre proksimaj unu al la alia kaj malkreskas monotone, dum la kurboj de la PID-klasanoj monotone kreskas. La tendencoj ankorau pli evidentiĝas, kiam oni kalkulas la mezumojn de N, G, A kaj P, I, D respektive kaj desegnas la koncernan diagramon (vidu fig. 2)

Se ni dispartigas la aron de la kazovicoj en ses subarojn tiamaniere, ke ĉiuj vicoj de unu subaro komenciĝas per la sama kazo, ni povas kalkuli la oftecon de la restantaj kvin kazoj. Ni interesiĝas do pri p_{ij} sub la kondiĉo, ke iu kazo k estu la plej ofta. Ni faris la kalkulon por k=1,2,4 (t. e. por nominativo, genitivo, akuzativo) kaj i=6 (prepozitivo). La kalkulo liveris la sekvantajn nombrojn:

	2	3	4	5	6	
NOM plej ofta:	5	8	14	22	51	(en %)
GEN plej ofta:	26	19	17	21	15	
AKU plej ofta:	15	17	20	27	22	

Se la kondiĉo, ke iu kazovico komenciĝas per k, ne influus la sinsekvon de la aliaj kazoj, ni povus atendi nur valorojn ĉirkau 20% en la supra tabelo. La kalkulo do montras, ke tia influo ekzistas: Se iu kazovico komenciĝas per N (nominativo), la sekvanta kazo tre verŝajne ne estos P (prepozitivo), sed male, kun kvindekprocenta verŝajneco oni povas profeti al P la lastan lokon. Tiuj ĉi faktoj fortikigas nian hipotezon, ke la kazooftecoj de unuopa substantivo dependas de ties semantika enhavo. La sekvanta diagramo ilustras la ĉi-supran tabelon:

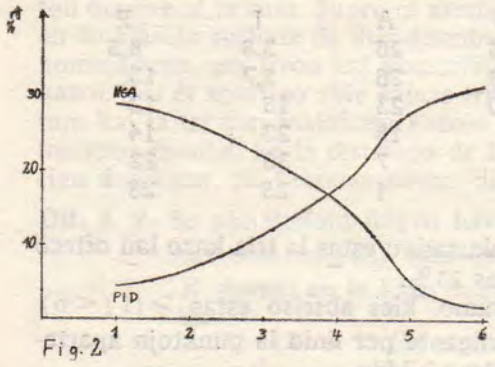


Fig. 2

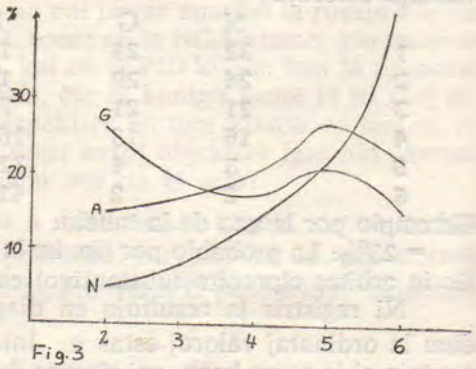


Fig. 3

Gis ĉi tie ni interesiĝis pri la deklinaciaj kazoj speciale en la rusa lingvo kaj komparis la statistikajn donitaĵojn kun tiuj el aliaj lingvoj. Plue ni konstatis la universalon de la ofteco de singularo kaj pluralo por kelkaj lingvoj. Ne esplorita estis tamen la demando, ĉu tia universaleco validas ankaŭ por unuopaj vortoj. Por kelkaj substantivoj ni faris la kalkulojn, uzante jenajn statistikajn vortarojn: [2] por la rusa, [7] por la hipsana, 19] por la germana kaj [10] por Esperento. Ni notas la relativan oftecon de singularo (SG).

Rusa [1]		Hispana		Germana		Esperanto	
GLAZ	9	ojo	10	Auge	32	okulo	9
VREM 4	97	tiempo	84	Zeit	90	tempo	79
JIZN 6	100	v.da	97	Leben	?	vivo	97
RABOTA	86	trabajo	74	Arbeit	?	laboro	?
MESTO	75	lugar	80	Ort	50	loko	57
DOM	76	casa	88	Haus	91	domo	73
GOLOVA	93	cabeza	87	Kopf	85	kapo	94
WKOLA	88	escuela	69	Schule	75	lernejo	?
GOROD	85	ciudad	80	Stadt	75	urbo	68
OKNO	65	ventana	67	Fenster	?	fenestro	?
STRANA	66	país	64	Land	77	lando	68
CAST 6	77	parte	85	Teil	71	parto	70
GOD	85	ano	28	Jahr	?	jaro	36
SLOVO	58	palabra	35	Wort	?	vorto	47
CAS	61	hora	48	Stunde	51	horo	59
DRUG	39	amigo	55	Freund	?	amiko	61
KNIGA	53	libro	63	Buch	64	libro	66
4 ZYK	87	lengua	76	Sprache	?	lingvo	77
RUKA	41	mano	56	Hand	68	mano	50
NOGA	26	pie	56	Fuß	53	piedo	17
DEN 6	73	dîa	57	Tag	?	tago	60

[1]) La rusaj vortoj estas transliterumitaj laŭ la sistemo priskribita en [11], nome

Я	Ь	Ш	Ч	Ж	Ю	Х	Ц
4	6	W	C	J	H	X	Q

Superrigardante tiun ĉi liston, oni konstatas, ke la devioj ĝenerale estas ne tro grandaj. Nur la eksterordinare granda ofteco de la singularo de GOD (»jaro«) frapas nian atenton. Ĝi estas kaŭzita de la aparta maniero, per kiu en la rusa lingvo oni citas jarnombrojn (»en la 1969-a jaro«).

3. Kazo-oftecoj en unuopaj semantikaj klasoj.

Post la diversaj komparoj de statistikaj donitaĵoj en pluraj lingvoj ni turnas nian atenton definitive al la studo de la distribuo de la ses rusaj kazoj, studante antaŭ ĉio la influon, kiun ekzercas la semantika enhavo de la substantivoj. Tiucele ni dispartigis la substantivojn en kelkajn klasojn, tiel ke la membroj de unu klaso havas komunajn trajtojn. La unua esplorota kaj plej granda klaso estas la personaloj, do la substantivoj, kiuj nomas personojn.

3. 1 Personaloj.

El 450 substantivoj ni trovis 103, kiuj apartenas al la supra klaso. Por ĉiu el ili ni kalkulis la kazo-oftecojn; jen kelkaj ekzemploj:

Rusa/Esperanta	vorto	N	G	D	A	I	P	kazovico
LHDI	homoj	43	31	8	8	7	3	NGADIP
REB4TA	infanoj	67	15	7	5	4	1	NGDAIP
BRAT	frato	66	12	10	7	4	2	NGDAIP
MAT6	patrino	60	14	8	11	7	1	NGADIP
MAL6CIK	knabo	58	23	5	8	4	1	NGADIP
PIONER	pioniro	41	39	8	5	6	0	NGDIAP
WKOL6NIK	lernanto	38	41	11	4	4	2	GNDAIP

Ni vidas, ke por ĉiu substantivo la procentaj oftecoj kaj la kazovico similas al la donitaĵoj de la aliaj vortoj. Tio fariĝas eĉ pli okulfrapa, kiam oni povas kompari kun la kvantoj ricevitaj por aliaj semantikaj klasoj.

Ni kalkulis la oftecon de ĉiu kazo en la personala klaso averaĝante la koncernajn oftecojn de la menciitaj 103 vortoj.

La mezumoj estas:

N=54,2%	A= 6,2%
G=22,2%	I= 9,0%
D= 7,1%	P= 1,2%

Por testo, ĉu tiuj oftecoj estas stabilaj, ni dividis la traktatajn vortojn en la tri grupojn de la 32 plej oftaj, 27 mezaj kaj 44 maloftaj substantivoj, kalkulante por ĉiu grupo la averaĝojn. Tie montriĝis nur malgrandaj devioj, kiuj ĉe la plej maloftaj kazoj estis iom pl girandaj ,ekz-e por dativo ni havis 6,7%—7,8%—6,4%. Por la 59 plej oftaj personaloj ni ricevis nominativan mezumon de 53,1% kaj genitivan de 21,4%.

La devio de la unuopaj konstatitaj kvantoj Xi (i=1,2,. . ., n) for de la averaĝo m estas nombre mezurebla per la varianco S^2 au la standarda devio S, kiu kalkuliĝas jene:

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - m)^2$$

Do S^2 estas la meza kvadrata diferenco.

Statistiko de la 59 plej oftaj substantivoj liveris standardajn deviojn de 12,9 por nominativo kaj 10,0 por genitivo. Tio signifas, ke normale la ofteco de nominativo situas inter 40% kaj 66%, dum la respondaj ciferoj por genitivo estas 11% kaj 32%. Fakte ekster tiuj intervaloj lokiĝas respektive 20 el 59 substantivoj, t. e. 34% — nombro, kiun oni teorie atendas. —

Komparante la ricevitajn averaĝojn kun la mezumoj de ĉiuj substantivoj, ni konstatas la sekvantajn apartecojn de la personaloj:

— nominativo estas multe pli ofta
— akuzativo estas multe pli malofta
— prepozitivo estas preskaŭ neuzata

kompare kun la mezumoj super ĉiuj substantivoj. La distribuoj de genitivo, dativo kaj instrumentalo nur malmulte diferencas de la ĝeneralaj distribuoj.

Konsiderante la averaĝojn oni atendas, ke la kazovico NGIDAP estas normala kaj tre ofta ĉe personaloj; fakte ni trovis 17 vortojn kun tiu ĉi kazovico, kiuj escepte de unu [1]) nomas personojn. 25 el 32 plej oftaj personaloj havas kazovicon komenciĝantan per NG kaj finiĝantan per P.

La oftecoj p_{ij} kalkulitaj por la personala klaso, montras denova tre ilustre, kiom la distribuo de la kazoj en tiu klaso diferencas de la ĝenerala mezumo. Jen la tabelo:

	N	G	D	A	I	P
1	94	5	0	0	1	0
2	5	77	3	6	9	0
3	0	12	27	23	37	1
4	0	3	37	33	25	2
5	1	2	28	38	28	2
6	0	0	2	2	2	94

Same kiel ni desegnis Fig. 1, ni konstruas la diagramon apartenantan al la supra tabelo.

Komparo kun Fig. 1 montras al ni, ke la kazo-distribuoj de la personaloj havas tute apartajn kvalitojn. Ĉi tie ekzistas altaj, pintaj maksimumoj kaj krutaj deklivoj ĉe nominativo, genitivo kaj prepozitivo, dum la tri aliaj kazoj sufiĉe uniforme distribuiĝas en la mezo — ĉiuj havas rondcifere 30% por $j=3,4,5$.

[1]) La escepto estas VETER = vento.

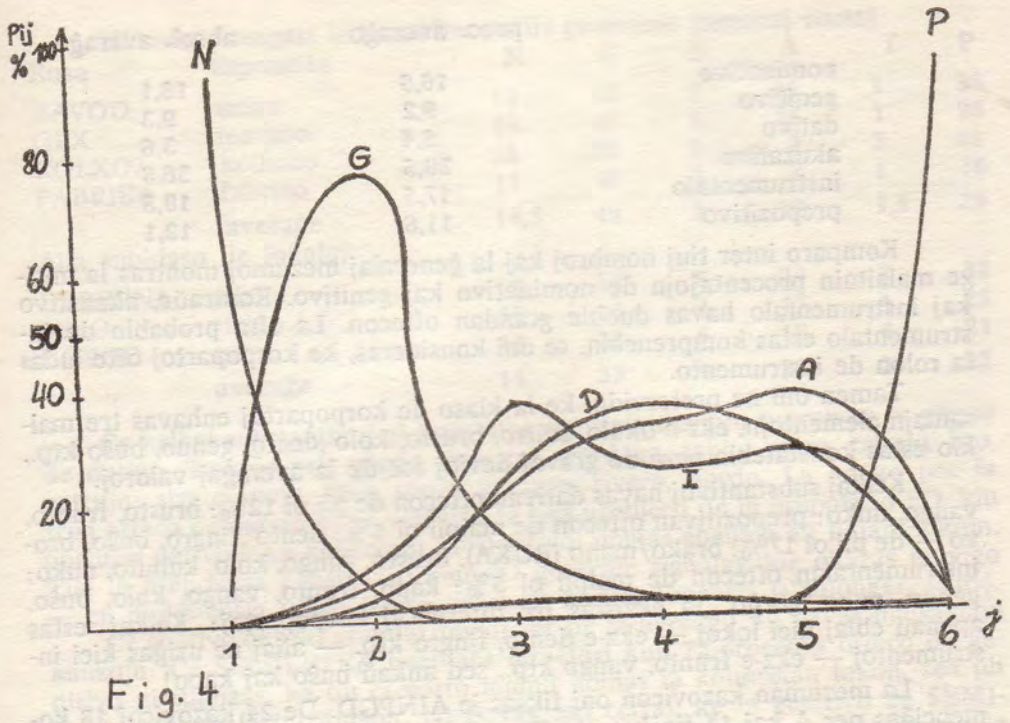


Fig. 4

3. 2 Animaloj.

Pruvinte statistikan rilaton inter la semantika distingilo »persono« kaj la kazoofteco, ni volas efektivigi la samajn kalkulojn por la kategorio »animalo«. Kvankam la nombro de la **substantivoj** ĉi tie esplorotaj estas malgranda, ĉar en [2] nur ok specimenoj estas listigitaj, ni tamen kredas, ke komparo kun la rezultoj de la personaloj donos profiton.

La traktataj animaloj estas: ĉevalo (dufoje), fiŝo, hundo, besto, bovino, kolombo, birdo. Ni mezumis lau du manieroj: La unuan fojon ni utiligis la procentajn ciferojn por averaĝado, la duan fojon la absolutajn nombrojn. La rezultoj estas

	proc. averaĝo	absol. averaĝo
nominativo	34	36
genitivo	30	29
dativo	4	4
akuzativo	19	19
instrumentalo	7	7
prepozitivo	5	5

Ni konstatas, ke ambaŭ metodoj de mezumado liveras preskaŭ samajn averaĝojn. La plej verŝajna kazovico estas NGAIPD. Kompare kun la personalaj mezumoj mirigas la alta probablo de akuzativo; korelative malpli ofta estas nominativo.

3. 3 Korpopartoj.

Ni averaĝis la unuopajn relativajn resp. absolutajn oftecojn de la kazoj de 24 substantivoj, kiuj nomas korpopartojn. La plej ofta substantivo okazis preskaŭ 750-foje, la plej malfota 20-foje. Rezultiĝis la sekvantaj valoroj:

	proc. averaĝo	absol. averaĝo
nominativo	16,5	18,1
genitivo	9,2	9,3
dativo	5,4	3,6
akuzativo	39,5	36,8
instrumentalo	17,5	19,8
prepozitivo	11,6	12,1

Komparo inter tiuj nombroj kaj la ĝeneralaj mezumoj montras la mirige malaltajn procentaĵojn de nominativo kaj genitivo. Kontraŭe, akuzativo kaj instrumentalo havas duoble grandan oftecon. La alta probablo de instrumentalo estas komprenebla, se oni konsideras, ke korpopartoj ofte ludas la rolon de instrumento.

Tamen oni ne pretervidu, ke la klaso de korpopartoj enhavas tre malsamajn elementojn, ekz-e okulo, ŝultro, brusto, kolo ,dento, genuo, buŝo ktp., kio estas konstatebla pere de gravaj devioj for de la averaĝaj valoroj.

Kelkaj substantivoj havas dativan oftecon de pli ol 12%: brusto, frunto, vango, nuko; prepozitivan oftecon de malpli ol 5%: dento ,fingro, buŝo, brovo — de pli ol 17%: brako/mano (RUKA), brusto, vango, kolo, kubuto, nuko; instrumentalan oftecon de malpli ol 5%: kapo, frunto, vango, kolo, buŝo, genuo, nuko, gorĝo. Ja ekzistas tre diferencaj korpopartoj: Kelkaj estas apenaŭ eblaj kiel lokoj — ekz-e dento, fingro ktp. — aliaj ne uziĝas kiel instrumentoj — ekz-e frunto, vango ktp., sed ankaŭ buŝo kaj kapo!

La mezuman kazovicon oni fiksas je AINPGD. De 24 kazovicoj 18 komenciĝas per A kaj 14 finiĝas per D — do, denove tute alia distribuo kompare kun la supre diskutitaj klasoj.

3. 4 Lokaloj.

La decido, ĉu donita substantivo estas klasifikebla kiel lokalo, ne ĉiam estas facila. Tial ni ne pritraktis dubajn specimenojn, ĉe kiuj oni ne estas certa pri ilia klasa aparteno. La plej oftaj elementoj de la lokala kategorio estas: loko, domo, flanko, lernejo, urbo, ĉambro ,pordo, tero, fenestro, tablo, strato ktp. Sume ni disponis pri 75 substantivoj, kiujn ni dividis en du klasojn: La unua klaso (31 membroj) entenas tiujn substantivojn, kiuj aperis ofte en la Ŝteinfeldtaj tekstoj kaj certe nomas lokojn; al la dua klaso apartenas 44 malpli oftaj substantivoj, kies lokaleco kelkfoje ne estas tiel certa. Kiel atendite, la mezumoj varias pli ol ĉe la personaloj:

	1-a klaso	2-a klaso
nominativo	11	15
genitivo	22	28
dativo	7,3	6,8
akuzativo	26	22
instrumentalo	5,2	4,6
prepozitivo	28,5	24

Tiuj ĉi ciferoj ectas mezumoj el la procentaj oftecoj liveritaj de la 31 resp. 44 substantivoj.

La mezuma kazovico por la unua klaso estas PAGNDI kaj por la dua GPANDI. Efektive ekzistas 6 substantivoj kun PAGNDI: angulo, kurso, laborĉambro, mondo, placo, mallumo — kaj 15 kun GPANDI: metiejo, konstruo, kilometro, metro, evoluo, dependeco, stacio, ĵurnalo ,lernejo, urbo, klaso, uzino, regiono, animo, mondo.

Ni volas kompari la kazooftecojn de proksime parencaj vortoj:

Rusa	Esperanto	N	G	D	A	I	P
ZAVOD	uzino	15	38	3	18	1	25
QEX	matiejo	14	41	3	15	1	26
KOLXOZ	kolĥozo	18	38	4	9	3	28
FABRIKA	fabriko	11	42	3	7	1	36
	averaĝe	14,5	40	3	12	1,5	29
Alia subklaso de lokaloj:							
DEREVN4	vilaĝo	3	32	6	24	2	32
SELO	vilaĝo	15	35	2	17	4	28
GOROD	urbo	15	38	5	18	4	21
	averaĝe	11	35	4	20	3	27

Ĉe kelkaj substantivoj ni konstatas signifajn deviojn supren en la ofteco de dativo. Traktas pri: pordo, fenestro, pordego, tablo, muro, tabulo, lito, patrujo; strato, vojo, linio spuro ŝtuparo, rivero, bordo. La kaŭzo por la troa ofteco de dativo estas verŝajne la tiea uzebleco de la prepozicio PO, kiu postulas dativon. La koncernaj substantivoj nomas ebenajn aŭ liniajn lokojn. Tial la prepozicia konstruaĵo kun PO ĝenerale signalas, kie okazas moviĝo — en dudimensia spaco (»sur«, »en«) aŭ en unudimensia (»laŭlonge«, »sur«).

Ĉe gentivo okazas grandaj devioj for de la averaĝo. Ili havas malsamajn kauzojn: Ĉe SEMILETKA (»sepjara lernejo«) kun 72-procenta ofteco de genitivo ni supozas, ke tiu ĉi vorto malpli nomas la konkretan lokon, sed pli la abstraktan, la institucion. Ni povas esti certaj, ke la genitivo de SEMILETKA malofte sekvas lokan prepozicion, sed ke ĝi estas uzata atribute. Ĉe aliaj substantivoj la granda genitiva ofteco estas kauzata de tio, ke kune kun la prepozicio U (»ĉe«) ili liveras lokansignifon.

Tio aplikiĝas ekz-e al VOROTA (»pordego«) kaj KOSTER (»lignofajro«), kiuj atingas 40% en genitivo.

	N	G	D	A	I	P
1	2,5	46	1	27	1	23
2	13	24	2	27	2	32
3	24	18	8	25	5	20
4	43	11	17	13	4	12
5	11	1	40	2	35	8
6	6	2	33	0	54	3

Tiujn ĉi valorojn ni ricevis per elnombrado de ĉiuj 75 lokaloj. La diagramo por p_{ij} esence diferencas de tiuj por personaloj kaj substantivoj ĝenerale. Por la 31 plej oftaj lokaloj oni tamen ricevas tre similan bildon. Por tiuj 31 ni kalkulis la standardajn deviojn por genitivo kaj prepozitivo, kiuj estas 11 resp. 10. Ili do havas la konatan grandecon.

Ĉe prepozitivo ekzistis 9 (=29%), ĉe genitivo 8 (=26%) specimenoj, kiuj deviis je pli ol 10 % resp. 11 % de la mezumoj. Ankaŭ tio estas kvantoj jam konataj de niaj antaŭaj kalkuloj. Por la p_{ij} -diagramo vidu Fig. 5

3. 5 Spacaj mezurunuoj.

La substantivoj por spacaj mezurunuoj multe distingiĝas lau kazoofteco de la nomoj por tempaj mezurunuoj. Tial ni ne enkondukis komunan klason, kies elementojn ni nomus »mezuraloj«. La kazooftecoj de la esploritaj kvin spacaj mezurunuoj permesas denovan duonigon de la klaso. Jen la donitaĵoj:

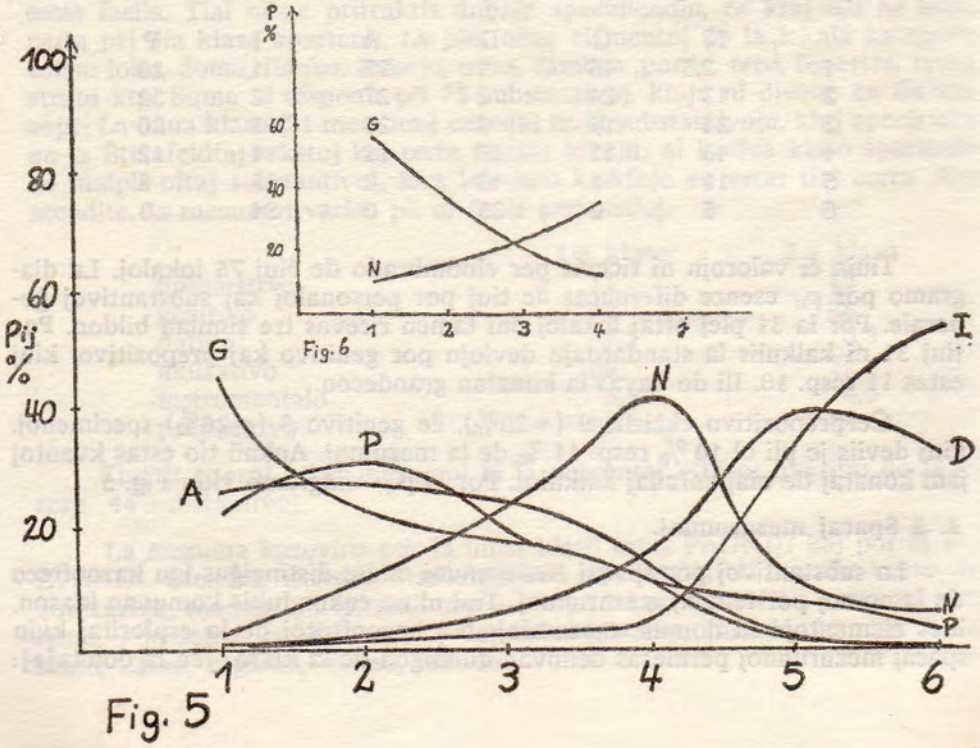
Ruse	Esperante	N	G	D	A	I	P
TONNA	tuno	5	86	0	6	2	2
KILOGRAMM	kilo	4	88	0	8	0	0
GEKTAR	hektaro	2	87	1	3	2	4
	averaĝe	4	87	0	6	1	2
METR	metro	3	80	1	7	1	8
KILOMETR	kilometro	4	73	3	9	0	14
	averaĝe	3,5	76,5	2	8	0,5	11

Por ambaŭ subklasoj la granda genitiva procentaĵo estas karakteriza, dum la ofteco de prepozitivo estas distinga por ili: La mezuroj de longo aperas pli ofte en prepozitivo ol la mezuroj de pezo kaj areo.

3. 6 Tempaj mezurunuoj.

La substantivojn de tiu ĉi klaso ni enklasigis en kvar subklasojn:

1-a klaso	N	G	D	A	I	P
sekundo	5	64	0	30	0	0
minuto	10	55	0	32	2	1
horo	12	57	4	25	1	2
monato	13	49	1	30	2	6
averaĝe	10	56	1	29	1	2
2-a klaso						
jaro	14	30	4	23	5	25
jarcento	20	43	0	15	4	19
3-a klaso						
tago	23	26	2	42	4	3
nokto	20	23	7	41	9	0
4-a klaso: sezonoj	N	G	D	A	I	P
printempo	28	15	8	17	31	2
somero	44	0	2	32	22	0



aŭtuno	32	24	4	16	20	4
vintro	38	11	2	36	11	2
averaĝe	36	12	4	25	21	2

En la unua subklaso troviĝas substantivoj havantaj tre altan procentaĵon por genitivo. Tiurilate ili similas al la spacaj mezurunuoj, distingiĝante de tiuj per la granda akuzativa ofteco. Tiu ĉi aparteco de la akuzativo klariĝas el tio, ke tiuj ĉi substantivoj uziĝas ne nur kiam daŭro estas komunikota — tiam aperus genitivo — sed ankau kiam tempopunkto estas sciigota. Ĉe la lokaloj uziĝas prepozitivo por indiki lokon, dum ĉe tiuj tempo-substantivoj la akuzativo estas postulata por simila tasko. La samo validas por la elementoj de la tria subklaso, kie genitivo ne plu havas troan oftecon: Tiuj ĉi substantivoj indikas pli ofte tempopunkton ol daŭron. Ankaŭ la subklaso de la sezonoj kondutas simile, sed diference de la aliaj ĉe ili la instrumentalo estas grava kazo. Ni trovas la kauzon en tio, ke la pura instrumentalo ĉi tie uziĝas por indiko de tempopunkto (»vintre«, »somere«). Laŭ nia opinio, tamen, tiuj ĉi instrumentalaj formoj estas rigidaj kaj devus esti nombrataj kiel adverboj. Ĉe la dua subklaso frapas la okulojn la granda ofteco de la prepozitivo; alie ol ĉe 'tago' au 'horo' oni uzas tie la sesan kazon por indiki tempopunkton, sed ne la akuzativon.

Interesa estas ĝenerala tendenco de la kvar subklasoj: Ju pli granda estas la ofteco de nominativo, des pli malgranda estas tiu de genitivo. Nia diagramo (Fig. 6, kun la klasoj numeroj sur la absciso) montras tion.

3. 7 Temposubstantivoj.

Krom la klaso de la tempaj mezurunuoj ni studis la aron de la substantivoj, kiuj nomas eventojn, okazaĵojn kaj epokojn. Grava kriterio por la klasifiko estis la demando, ĉu la donita vorto uziĝas kun la prepozicioj »dum«, »antaŭ«, »post« kaj »ek de«. Laŭ tiu ĉi difino la klaso de la tempo-substantivoj enhavas la vortojn »koncerto«, »teatro«, »laboro«, »milito«, «revolucio», »ludo«, »inflanaĝo«, »batalo«, »kunveno« ktp., sed la klasifikado ne estas tiel simpla kiel ĉe la personaloj. Ni utiligis la donitaĵojn de 15 plej oftaj tempo-substantivoj por kalkuli la mezumajn oftecojn. Ili estas

N=22%	A=20%
G=32%	I= 4,5%
D= 3%	P=19%

Komparante tiujn ĉi valorojn kun tiuj, kiujn ni ricevis por la lokaloj, ni konstatas grandan similecon. La mezuma kazovico estas GNAPID; ekzistas sep substantivoj kun tiu vico, nome: gazeto, teatro, konsilantaro, partio, komsomolo, armeo, unio — do, neniel tempo-substantivoj. Komparo de la relativaj oftecoj de tiuj sep substantivoj montras, ke ili apartenas al du specoj. Por »teatro« kaj »gazeto« ni havas la oftecojn 23, 34, 2, 21, 3, 17, kio preskaŭ koincidas kun la mezumoj de la temposubstantivoj, dum la restantaj kvin vortoj multe devias de tiu averaĝo.

3. 8 Kolektivoj.

Iom pli grandan klason formis la substantivoj nomantaj unuiĝojn (precipe de homoj) au ion tian; ni nomis ilin »kolektivoj«. Estis trovitaj 19 vortoj de tiu speco en la studita vortaro [2]. Montriĝas, ankaŭ ĉi tie la kazooftecoj distribuiĝas tute alimaniere ol ĉe la antaŭe esploritaj grupoj. Ni konstatis la sekvantajn averaĝojn:

N=21%	A= 9%
G=50%	I= 5%
D= 4,5%	P=10%

La grupo enhavas ekz-e jenajn elementojn: partio, komsomolo, armeo, unio, kolektivo, popolo, registaro, organizaĵo, komitato, brigado, familio, klubo.

La mezumaj kazovicoj estas tial GNPAID, GNAPID, GNPADI kaj GNAPDI, al kiuj apartenas la sekvantaj substantivoj: tempolimo, revolucio, domo, okupo, ludo, interrilato ,gazeto, teatro, konsilantaro, partio, komsomolo, armeo, unio, ekonomio, brigado, arto, ekipo (personara), jarcento.

Ni ekkonas, ke la scio de la kazovico de iu substantivo ne sufiĉas por semantika klasifikado, ĉar ofte — kiel oni vidas supre — al unu kazovico apartenas tre malsamaj vortoj. Tial nia tasko estas serĉi eblecojn por taksi la apartenon de iu substantivo al iu klaso. Se ekzistus sufiĉe sekura metodo por difini pere de kazooftecoj certan parencecon inter la substantivoj, tiam oni povus decidi, ĉu »patrujo« estu profere konsiderata kiel lokalo aŭ kolektivo. Kiujn eblecojn ni trovis, estos diskutite en la posta ĉapitro.

4. Parenco-distancoj.

Interesas nin la demando, ĉu estas eble konkludi el la ofteco de la kazoj de iu substantivo ĝian apartenon al unu el la supre difinitaj klasoj. Ni jam konstatis, ke la populacio de iu klaso posedas komunajn kvalitojn, kiuj ilustre montriĝas per malsamaj mezumoj kaj distribuoj en la p_{ij} — diagramo. La oftecajn mezumojn de iu klaso ni nun konsideras karakterizaj por ĝia populacio, dum la valoroj liverataj de unuopaj substantivoj estas rigardataj kiel statistike kaj semantike kaŭzataj devioj ĉirkaŭ la mezumoj.

Nia unue celo estas nun montri, ke lokaloj kaj personaloj povas esti tute klare separitaj. Por personaloj la mezuma kazovico estas NGIDAP, por lokaloj PAGNDI au GPANDI. La plej oftaj kazoj estas do nominativo, genitivo kaj instrumentalo unuflanke kaj genitivo, akuzativo kaj pepozitivo aliflanke. Ni volas loki personalojn kaj lokalojn en t.n. GAP-NGI-diagramon Al ĉiu substantivo ni aldifinas punkton en la x-y-ebeno, fiksante ĝian abscisan vahoron per la sumo de ĝiaj genitiva, akuzativa kaj prepozitiva oftecoj, dum la ordinato kalkuliĝas kiel la sumo de N, G kaj I. Oni povas ekspekti la sekvantajn kvantojn:

	personaloj	lokaloj
x=GAP	30%	74%
y=NGI	85%	43%

Ni taksas, ke en la unua kvadranto de la ebeno ekzistas por la du klasoj po unu regiono kaj ke tiuj regionoj ne intersekcas. La personaloj devas lokiĝi maldekstre alte, la lokaloj dekstre malalte en la diagramo. La pezocentro de la personala regiono estas (30, 85), tiu de la lokaloj estas (74, 43). La distribuon de la substantivoj en tia diagramo montras fig. 7.

Se oni lokas en la saman diagramon ankaŭ la aliajn substantivojn, tiam montriĝas, ke ili same grupiĝas ĉirkau la pezocentro de sia klaso. Kiam du regionoj en la GAP-NGI-diagramo havas ne-malplenan intersekcon, oni povas provi disigi ilin en alia ebeno. Ekz-e la substantivoj por korpopartoj kaj lokoj posedas en la GAP-NGI-ebeno grandan komunan regionon, dum en la ANI-GAP-ebeno ili nur tuŝiĝas.

Cele al apartigo de du klasoj ni elektis kiel abscison resp. ordinaton la sumojn de iliaj tri plej oftaj kazoj, ĉar tiuj situas lau nia sperto supre de 70 % kaj atribuas al la klasoj tre diferencajn regionojn en la ebeno, kondiĉe ke ambaŭ kazoparoj estu laŭeble malsamaj. Kiam tamen la tri plej oftaj kazoj de la du klasoj estas samaj, la metodo ne promesas sukceson. Tiam oni povus utiligi la donitaĵojn por tiuj kazoj, kie la plej grandaj diferencoj estas atendeblaj.

Jen la aspekto de la GAP-NGI-ebeno:

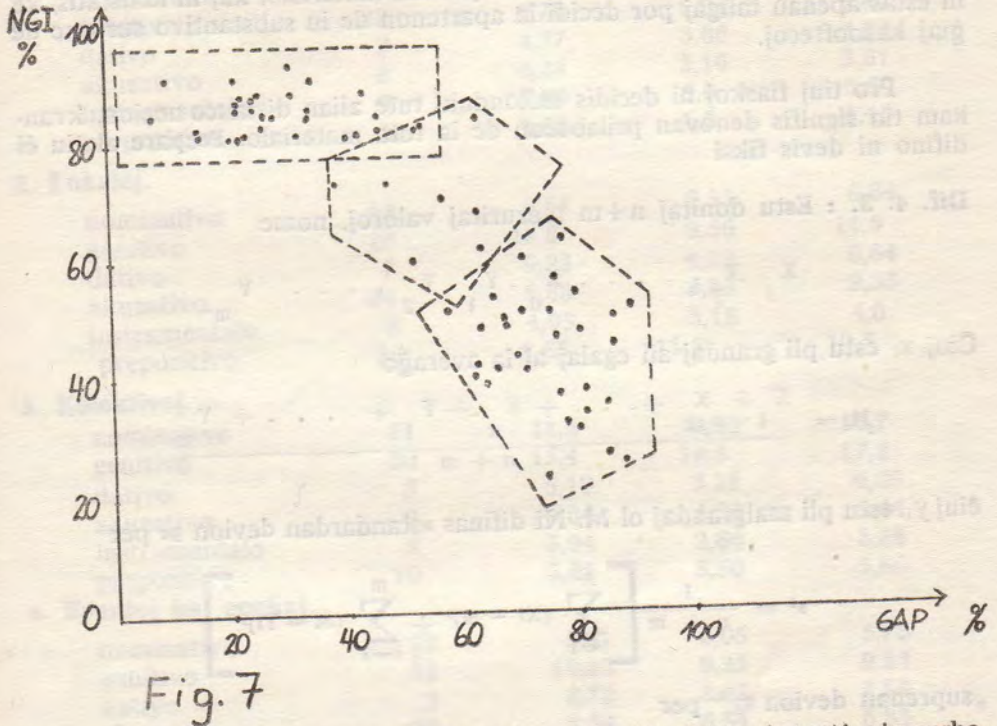


Fig. 7

Se ni transdonas al iu automato la taskon klasifiki substantivojn surbaze de iliaj kazooftecoj, ĝi povus solvi la problemon per la decido, en kiun regionon falas la esplorata vorto. Alia niaopinie pli eleganta solvo estus la difino de distanco de la substantivo for de klasa pezocentro. Tie sin ofertas tuj la konata distanco en la sesdimensia eŭklida spaco:

Dif. 4. 1: La distanco d_1 de iu substantivo L kun la kazooftecoj N, G, D, A, I, P for de iu pezocentro de la klaso K kun la mezumoj $N_k, G_k, D_k, A_k, I_k, P_k$, estu

$$d_1(L, K) = \sqrt{(N - N_k)^2 + (G - G_k)^2 + \cdots (P - P_k)^2}$$

Por niaj celoj tamen la supra difino ne estas taŭga. Supozu $N = 50\,\%$, $p_k = 5\%$, $N = 60\,\%$, $P = 15\,\%$, tiam la diferencoj $N - N_k$ kaj $P - P_k$ (ambau $= 10\%$) havus la saman pezon en la kalkulo de la distanco, kvankam objektive la devio de P_k for de P ŝajnas al ni multe pli grava, ol tiu de N_k for de N. Tial ni kredis devi relativigi tiujn ĉi deviojn:

Dif. 4. 2: La distanco d_2 inter L kaj la klaso K estu

$$d_2(L, K) = \sqrt{\left(\frac{N - N_K}{N}\right)^2 + \ldots + \left(\frac{P - P_K}{P}\right)^2}$$

Ni elprovis tiun distanco-nocion kaj trovis ĝin ankaŭ sufiĉe maltaŭga. Se ekz-e N estas granda nombro, la divido per N tro malgrandigas deviojn, se ĝi estas malgranda, la rezulto esos trograndigita distanco. Ankau aliajn distanco-difinojn ni studis kaj elprovis per nia materialo, kaj ni konstatis, ke ili estas apenaŭ taŭgaj por decidi la apartenon de iu substantivo surbaze de ĝiaj kazooftecoj.

Pro tiuj fiaskoj ni decidis enkonduki tute alian distanco-nocion, kvankam tio signifis denovan prilaboron de la tuta materialo. Prepare al tiu ĉi difino ni devis fiksi

Dif. 4. 3. : Estu donitaj n+m mezuritaj valoroj, nome

$$X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots Y_m.$$

Ĉiuj x_i estu pli grandaj aŭ egalaj al la averaĝo

$$M = \frac{X_1 + X_2 + \ldots + X_n + Y_1 + \ldots + Y_m}{n + m}$$

ĉiuj y_i estu pli malgrandaj ol M. Ni difinas »standardan devion s« per

$$s^2 = \frac{1}{n + m}\left[\sum_{i=1}^{n} (Xi - m)^2 + \sum_{i=1}^{m} (M - Yi)^2\right]$$

„suprenan devion s_0 " per

$$s_0^2 = \frac{1}{n}\sum_{i=1}^{n} (Xi - M)^2$$

kaj »malsuprenan devion s_1« per

$$s_1^2 = \frac{1}{m}\sum_{i=1}^{m} (M - Yi)^2$$

La standardan devion ni jam renkontis en ĉapitro 3; ĝi estas mezuro de la devio ĉirkau la averaĝo. La suprena devio mezuras, kiom la mezuritaj valoroj devias supren for de la averaĝo, la malsuprena devio montras la devion malsupren. Ĉe simetrie distribuitaj kvantoj oni atendas, ke s_0 kaj s_1 estas preskaŭ samaj. La afero statas alie ĉe niaj problemoj, ĉar tie ne ekzistas simetriaj distribuoj de la procentaj valoroj, kuŝantaj en fermita intervalo. Ju pli proksime al intervala limo kuŝas M, des pli grandaj estos la diferencoj inter s^0 kaj s_1.

Por la kvin sufiĉe frekventaj klasoj de personaloj, lokaloj, kolektivoj, temposubstantivoj kaj korpopartoj ni kalkulis la averaĝojn de la ofteco de la kazoj kaj la tri deviojn. Por la animaloj ni povis fiksi nur la averaĝojn kaj la standardajn deviojn, ĉar ni havas tro malmulte da specimenoj en tiu ĉi klaso por povi doni iom sekurajn valorojn.

Jen tabelo de la averaĝoj kaj devioj:

kazo	averaĝo	s_o	s_1	s
1. Personaloj.				
nominativo	54	12,5	12,8	12,65
genitivo	22	9,78	9,04	9,40
dativo	7	4,77	3,66	4,22
akuzativo	6	4,28	3,16	3,81
instrumentalo	9	7,80	5,0	6,40
prepozitivo	1	1,93	1,0	1,45
2. Lokaloj.				
nominativo	13	6,84	6,35	6,64
genitivo	25	12,8	9,56	11,5
dativo	7	9,23	4,22	6,64
akuzativo	24	9,88	8,86	9,35
instrumentalo	5	4,95	3,18	4,0
prepozitivo	26	8,65	11,5	10,5
3. Kolektivoj.				
nominativo	21	11,5	9,93	10,7
genitivo	50	15,4	19,5	17,2
dativo	5	5,19	3,28	4,09
akuzativo	9	4,64	4,20	4,44
instrumentalo	5	3,94	2,66	3,26
prepozitivo	10	6,01	5,50	5,80
4. Eventoj kaj epokoj.				
nominativo	22	5,4	6,05	5,76
genitivo	32	10,25	9,35	9,85
dativo	3	2,72	2,42	2,59
akuzativo	20	5,34	6,54	5,85
instrumentalo	5	3,06	3,05	3,06
prepozitivo	19	7,30	7,60	7,40
5. Korpopartoj.				
nominativo	17	10,0	8,84	9,40
genitivo	9	5,04	4,80	4,88
dativo	5	4,65	3,27	4,02
akuzativo	40	16,1	9,16	12,5
instrumentalo	18	12,8	12,2	12,5
prepozitivo	12	7,45	5,65	6,5
6. Animaloj.				
nominativo	34			8,7
genitivo	30			11,0
dativo	4			2,24
akuzativo	19			8,3
instrumentalo	7			3,6
prepozitivo	5			4,1

Superrigardante tiujn ciferojn ni ekkonas, ke iu klaso estas karakterizata ne sole per la meza ofteco de la kazoj, sed ankaŭ per la devioj — standarda, suprena kaj malsuprena — kiuj indikas ,kiaj devioj ĉe donita specimeno estas allaseblaj, por ke ĝi povu esti konsiderata kiel membro de tiu klaso. Tiaj devioj do estas speco de distanco au mezuro por la parenceco de iu vorto rilate al iu klaso.

Por povi formule difini tian distancon, ni enkondukas unue la sekvantajn signojn: Estu donita substantivo L, kies kazooftecoj estu (x^1; x^2;... ; x^6). La averaĝaj oftecoj de iu klaso K estu (m^1; m^2; ... m^6), la suprenaj kaj malsuprenaj devioj estu

$$(s_{01}, s_{11} : s_{02}, s_{12} : \ldots : s_{06}, s_{16})$$

Dif. 4. 4: La distanco inter L kaj la klaso K estu

$$D_K(L) = \sum_{i=1}^{6} \left(\frac{mi - Xi}{si} \right)^2$$

kie oni metu $s_i = s_{0i}$, kiam x_i siperas m_i kaj alikaze $s_i = s_{1i}$

Surbaze de la difino de la suprena kaj malsuprena devioj oni taksas, ke la substantivoj apartenantaj al K havas distancon de ĉirkau 6 for de K. Se ĉe iu substantivo la distanco al K estas pli granda, tiam oni povas supozi la ekziston de iu aparteco, kies klarigo eble estas interesa.

Por kelkaj personaloj ni kalkulis la distancon for de la koncerna klaso. Jen la rezultoj:

substantivo L		D_{per} (L)	subsantivo (L)		D_{per} (L)
LHDI	homoj	3,06	REB4TA	infanoj	2,74
substantivo L		Dper(L)	substantivo L		Dper(L)
TOVARI5	kamarado	2,88	BRAT	frato	3,86
DEVUWKA	knabino	2,08	DEVOCKA	knabino	3,20
MAT6	patrino	2,57	CELOVEK	viro	6,13
MAL6CIK	knabo	1,63	DETI	infanoj	5,85
MAMA	panjo	8,32	NACAL6NIK	ĉefo	3,88

Ni ekvidas, ke la distancoj ĝenerale eĉ malsuperas 6. Sed povi taksi tiun ĉi fakton, ni devas scii, ĉu substantivoj el aliaj klasoj havas aliajn distancojn for de la personala klaso.

Se tiuj distancoj estus simile grandaj, tiam la distanco ne estus mezuro de parenceco.

Tial ni kalkulas por la averaĝo de ĉiu klaso la distancon al la restantaj klasoj:

L	D_{per}	D_{lok}	D_{kol}	D_{okaz}	D_{korp}
personalo	0	44,7	14,7	50,2	38,6
lokalo	197	0	20,0	6,41	18,1
kolektivo	37,9	9,9	0	7,87	78,9
okazaĵoj	108	3,5	9,2	0	28,2
korpopartoj	108	15,0	60,3	39,2	0
animaloj	18,1	13,6	7,9	8,7	30,4

La tabelo montras, kiel granda estas la danĝero, ke aŭtomato laboranta laŭ la distanco-metodo malĝuste klasifikas donitan substantivon. Ju pli grandaj estas tiuj supraj distancoj, des pli malgranda estas la probablo, ke substantivo estos klasifikita malĝuste laŭ siaj kazooftecoj. Ni konkludas jenon:

Personaloj verŝajne estos klasifikitaj ĝuste; ĉi tie ekzistas malgranda danĝero kolizii kun la kolektiva kategorio. La kolektivoj povus hazarde fali en la aron de la lokoj aŭ okazaĵoj. Malfacila estas la apartigo de lokoj kaj okazaĵoj: La lastaj distancas je 3,5 de la lokala klaso, dum la unuaj havas distancon de 6,41 de la klaso de la okazaĵoj. Substantivoj nomantaj korpopartojn povas esti konfuzitaj kun lokaloj, dum la animaloj havas ŝancojn esti erare klasifikitaj kiel kolektivoj.

Por dudeko da substantivoj ni elkalkulis la distancojn for de la kvin supraj klasoj. Jen la tabelo:

VAGON	vagono	155	6,9	11,3	4,4	47,8
RODINA	patrujo	31	5,4	3,9	16,6	45,6
NEBO	ĉielo	132	1,0	6,2	4,0	16,1
VOZDUH	aero	259	6,9	22,0	8,3	18,9
DETI	infanoj	5,9	20,4	6,2	20,4	50,0
KOMSOMOLEQ	komsomolano	0,9	44,6	11,6	42,0	44,3
RUKOVODITEL6	direktoro	2,3	40,0	19,2	48,3	40,6
SMENA	ŝanĝo/laborvico	99	10,3	67	29,1	16,9
PARTI4	partio	38,2	9,2	1,5	9,9	94,1
IGRA	ludo	33,9	10,6	7,8	5,5	19,3
GLAZ	okulo	75,3	25,1	56,1	45,7	3,4
LOWAD6	ĉevalo	49,0	10,3	13,9	9,2	10,3
RYBA	fiŝo	55,6	12,3	36,1	14,1	22,3
SOBAKA	hundo	44,7	—	—	—	— [1]
JIVOTNOE	besto	42,5	14,4	3,4	—	—
KOROVA	bovino	21,8	—	—	—	—
KON6	ĉevalo	44,3	—	5,1	—	—
GOLUB6	kolombo	18,0	—	—	—	—
PTIQA	birdo	3,6	—	8,1	—	—

[1] La mankantaj nombroj estas tre grandaj kaj seinteresaj.

Tiujn ĉi substantivojn ni elektis per hazardaj nombroj (krom la animaloj). Ni konstatas, ke ili tre ofte estas ĝuste klasifikitaj, kvankam kelkaj neatenditaj okazis: Vagono estis pli verŝajne okazaĵo ol loko, kaj »patrujo« parencas pli al la kolektivoj ol al la lokaloj. Pripensinte tion, oni ne scias, ĉu tiu kvalifiko de »patrujo« estas malĝusta!

VOZDUX (aero) ne estis uzita por la kalkulo de la mezumoj kaj devioj, ĉar ni ne sciis kun certeco, kien ĝin meti. Nun montriĝas, ke »aero« fakte estas iom problema, ĉar ĝia plej malgranda distanco estas 6,9 — kion ni konsideras sufiĉe granda — per kio ĝi estas plej parenca al la lokaloj. DETI (infanoj) estas ĝuste ekkonita kiel personalo, sed ĝi povus esti ankaŭ kolektivo kun preskaŭ sama probablo. Tio efektive havas sencon, ĉar DETI estas substantivo, kiu aperas en tiu formo nur kiel plurado. Ĉe SMENA ni povas esti certaj ke ĝi ne apartenas al unu el niaj klasoj. La kaŭzo estas tre verŝajne la plursignifeco. Laŭ [12] SMENA signifas: ŝanĝ (ad) o, alterno, laborvico, alternularo kompleto, anstataŭantoj, anstataŭanta generacio! La bestonomoj tre ofte akordiĝis kun neniu klaso, kelkfoje kun kolektivoj aŭ personaloj.

5. Postparolo.

La antaŭe raportitaj studoj pruvas interrilaton inter la signifo de substantivo kaj ties kazoofteco, t.e. inter semantika enhavo — aŭ aparteno al semantika klaso — kaj probableca distribuo de tiu substantivo. Eblas difini nocion de distanco inter substantivo kaj semantika klaso, kiu estas mezuro de parenceco aŭ verŝajno de aparteno al tiu klaso.

Se oni havus aŭtomaton, kapablan pri sintaksa analizo de la rusa lingvo, ĝi povus memstare (kun certa erarprobablo) klasifiki la oftajn substantivojn de la analizitaj de ĝi tekstoj. Tia klasifikado povas esti deirpunkto por plibonigado de la analiza algoritmo.

Fine ni rimarkigas, ke la rusajn vortojn ni tradukis laŭ [12] kaj krome uzis kontrole [13] kaj [14]. La matematikajn fakterminojn ni prenis el [15]. Oni pardonu, se niaj fakaj neologismoj »personalo«, »lokalo« kaj »animalo« kiel mallongigoj de tro longaj parafrazoj kiel »substantivo nomanta personon« ktp. ne plaĉas. Ni kredas, ke ni ŝparis per tio multan lokon kaj tempon!

6. Bibliografio.

[1] H. D. Maas, Die Synthese deutscher Sätze im Zusammenhang mit maschineller Sprachübersetzung. In: Muttersprache, 1969, Heft 9/10.
[2] E. Steinfeldt, Russian Word Count, Moscow (sen jaro).
[3] W. Fucks, Nach allen Regelen der Kunst, Stuttgart 1968.
[4] A. W. Thomson, Vortofteco. En: Scienca Revuo, N-ro 74, 1968.
[5] H. D. Maas, Homographie und maschinelle Sprachübersetzung. En: Linguistsche Arbeiten des Germanistischen Instituts und des Instituts für Angewandte Mathematik der Universitäat des Saarlandes (mallongigo: LA), Nr. 8, 1969.
[6] G. K. Zipf, The Psycho-Biology of Language. Cambridge (USA), 1965.
[7] A. Juilland E. Chang-Rodriguez, Frequency Dictionary of Spanish Words, The Hague, 1964.
[8] H. Meier, Deutsche Sprachstatistik. Hildesheim 1964.
[9] W. Kaeding, Häufigkeitswörterbuch der deutschen Sprache. Steglitz bei Berlin, 1898.
[10] L. Blaas, Statistiko de 50.000 tekstvortoj. En: Esperantologio, Copenhagen, Nr. 2 (1950), Nr. 3 (1951).
[11] Janda /Rothkegel/ Zimmermann: Dokumentation eines Programms zur Analyse russischer Sätze. En: LA 7, 1969 (klarigo de LA vidu sub [5]).
[12] E. A. Bokarev, Rusa-Esperanta Vortaro, Moskvo 1966.
[13] E. —D. Krause, Esperanto-Deutsches Wörterbuch, Leipzig 1967.
[14] H. Wingen, Wörterbuch Deutsch-Esperanto, Limburg (Lahn 1954.
[15] C. M. Bean, Matematika Terminaro, Heronsgate (Angl.) 1954.